JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. XII No 1, Desember 2025, hlm. 175 – 182 ISSN 2550-0201 (Online)
DOI: http://dx.doi.org/ 10.33330/jurteksi.v12i1.4355
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

EFFICIENTNET MODEL FOR BONE AGE PREDICTION

Elliya Sestri[1*], Widi Hastomo [1], Silvia Ningsih [1]
[1]Information Technology, Institut Teknologi dan Bisnis Ahmad Dahlan
email: *ellyasestri24@gmail.com

Abstract: Accurate bone age estimation is essential for monitoring pediatric growth, diagnosing endocrine disorders, and supporting clinical decision-making. Although deep learning has improved prediction accuracy, limited studies have systematically examined how increasing model depth affects performance and reliability. This study evaluates the effectiveness of progressively deeper convolutional neural networks, specifically EfficientNet variants B0 to B5, for bone age estimation from hand radiographs. Experiments were conducted using 12,611 hand X-ray images from the RSNA Pediatric Bone Age Challenge dataset on Kaggle. To ensure fair comparison, all models were trained using a unified and consistent training pipeline. Model performance was evaluated using Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Concordance Correlation Coefficient (CCC), and Pearson correlation coefficient. The results show a consistent improvement in prediction accuracy as model depth increases. Among the evaluated models, EfficientNet-B5 achieved the best performance, with an MAE of 21.5 months, MAPE of 6.23%, CCC of 0.9148, and Pearson's r of 0.9203. These findings confirm that model scaling plays a critical role in enhancing prediction robustness and clinical reliability. Future work should emphasize external validation across diverse populations and incorporate interpretability techniques, such as Grad-CAM, to improve clinical transparency and trust.

Keywords: bone age prediction; deep learning; model evaluation; clinical validation

Abstrak: Estimasi usia tulang yang akurat sangat penting untuk memantau pertumbuhan anak, mendiagnosis gangguan endokrin, dan mendukung pengambilan keputusan klinis. Meskipun pembelajaran mendalam telah meningkatkan akurasi prediksi, studi yang secara sistematis meneliti bagaimana peningkatan kedalaman model memengaruhi kinerja dan keandalan masih terbatas. Studi ini mengevaluasi efektivitas jaringan saraf konvolusional yang semakin dalam, khususnya varian EfficientNet B0 hingga B5, untuk estimasi usia tulang dari radiografi tangan. Eksperimen dilakukan menggunakan 12.611 gambar sinar-X tangan dari dataset RSNA Pediatric Bone Age Challenge di Kaggle. Untuk memastikan perbandingan yang adil, semua model dilatih menggunakan alur pelatihan yang terpadu dan konsisten. Kinerja model dievaluasi menggunakan Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Concordance Correlation Coefficient (CCC), dan koefisien korelasi Pearson. Hasil menunjukkan peningkatan yang konsisten dalam akurasi prediksi seiring dengan peningkatan kedalaman model. Di antara model yang dievaluasi, EfficientNet-B5 mencapai kinerja terbaik, dengan MAE sebesar 21,5 bulan, MAPE sebesar 6,23%, CCC sebesar 0,9148, dan Pearson's r sebesar 0,9203. Temuan ini menegaskan bahwa penskalaan model memainkan peran penting dalam meningkatkan optimasi prediksi dan keandalan klinis. Penelitian selanjutnya dapat menekankan validasi eksternal di berbagai populasi dan menggabungkan teknik interpretasi, seperti Grad-CAM, untuk meningkatkan transparansi dan kepercayaan klinis.

Kata kunci: prediksi usia tulang; deep learning; evaluasi model; validasi klinis

INTRODUCTION

Bone age is an important biological indicator in evaluating growth and development in children and adolescents[1]. Clinically, traditional bone age assessments, such as the Greulich–Pyle and Tanner–Whitehouse methods, rely on visual interpretation of radiographs by radiologists, which is often time-consuming and prone to inter-observer variation[2]. Advances in deep learning technology offer opportunities to automate this process, improving objectivity and diagnostic efficiency[3]. The implementation of convolutional neural network (CNN) models in bone age estimation tasks has been shown to capture complex anatomical features in radiographic images that are difficult to identify manually, thus supporting higher prediction accuracy compared to manual methods.

Despite promising findings regarding EfficientNet's potential, a comprehensive comparative understanding of different scaled variants (B0 through B5) for bone age prediction remains limited[4]-[7]. Existing studies have predominantly focused on specific individual variants, such as B3, without systematically examining the trade-offs between accuracy, computational complexity, and resource requirements across the full spectrum of available models[8], [9], [10]. Furthermore, the evaluation of clinically meaningful agreement metrics, such as the Concordance Correlation Coefficient (CCC), has often not been prioritized in model assessment[11].

Therefore, this study aims to: (1) evaluate and compare the performance of six EfficientNet architecture variants (B0 to B5) in estimating bone age from hand radiographs, considering both standard accuracy metrics (e.g., MAE, RMSE) and clinical agreement metrics (CCC, Pearson’s R); (2) analyze the computational efficiency and parameter count of each variant to identify the optimal model in terms of performance-practicality balance; and (3) compare the best-performing EfficientNet model against other widely-used CNN architectures, namely ResNet50 and DenseNet121, as baseline references. The outcomes of this research are expected to offer practical, evidence-based guidance in selecting an effective and efficient deep learning architecture for automated bone age assessment, particularly in settings with varying computational resources.

METHOD

This study employs a quantitative experimental approach utilizing an EfficientNet-based deep learning architecture for bone age prediction from pediatric hand radiographs. The analysis was conducted on the publicly available RSNA Bone Age dataset [19], which comprises 12,611 left-hand radiographic images - the standard anatomical region for bone age assessment. Each image is annotated with bone age (in months) as determined by radiologists, chronological age (in months), and patient gender (male/female). The dataset spans the complete pediatric growth period, with chronological ages ranging from 1 to 228 months (approximately 19 years), and corresponding bone age annotations covering a similar developmental spectrum.

All radiographic images underwent standardized preprocessing, including pixel intensity normalization and resizing to EfficientNet input dimensions (224×224 pixels for B0 variant, with proportional scaling for B1-B5 architectures). Data augmentation techniques -

including rotation, horizontal flipping, and contrast adjustment - were applied to enhance dataset diversity and mitigate overfitting[20], [21].

We systematically evaluated EfficientNet variants B0 through B5 to examine how architectural scaling across depth, width, and input resolution dimensions influences predictive accuracy and computational efficiency. Models were initialized with ImageNet pretrained weights and fine-tuned on the radiographic dataset, with final layers modified to output a single continuous bone age value (in years). Training employed the Adam optimizer with a learning rate of 1e-4, batch size of 16, and early stopping based on validation performance. Model efficacy was quantified using multiple metrics: Mean Absolute Error (MAE), Mean Squared Error (MSE) [22], Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE) [23], Symmetric MAPE (sMAPE), Concordance Correlation Coefficient (CCC), and Pearson Correlation (Pearson R), comparing predictions against radiologist-assigned bone age references. EfficientNet performance was benchmarked against ResNet50 and DenseNet121 architectures, with all experiments implemented in TensorFlow/Keras using GPU acceleration and validated through k-fold cross-validation to ensure generalizability.

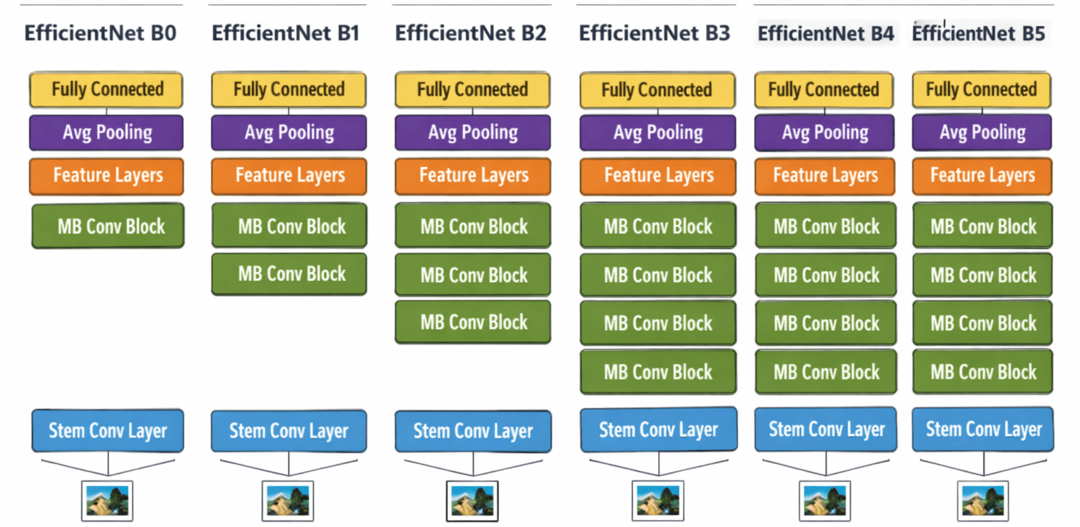


Image 1. EfficientNet B0-B5

EfficientNet uses a compound scaling approach to proportionally balance image depth, width, and resolution. The formula (compound scaling) is:

$$\text{depth: } d = \alpha^{\phi}, \text{width: } w = \beta^{\phi}, \text{resolution: } r = \gamma^{\phi} \quad (1)$$

with the provision of:

$$\alpha \cdot \beta^2 \cdot \gamma^2 \approx 2 \text{dan} \alpha, \beta, \gamma > 1 \quad (2)$$

ϕ is the compound scaling coefficient for the model version (B0 - B5). α, β, γ are constants chosen to balance depth, width, and resolution. Each successive model (B1–B5) increases ϕ, resulting in a larger, more capable network.

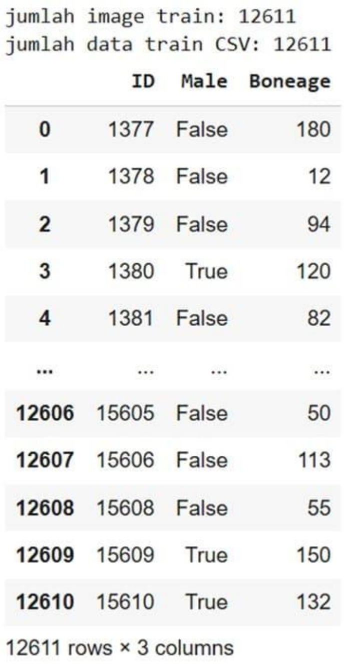
EfficientNet uses a clever approach called *compound scaling*, which grows the network in three dimensions at once: depth (more layers), width (more channels per layer), and input image resolution. Instead of arbitrarily making the network bigger, the creators found an optimal balance so that each increase in size improves performance efficiently.

Think of it like upgrading a camera: instead of only increasing the megapixels (width) or adding more lenses (depth), EfficientNet adjusts all aspects together including the size of the image it looks at so that every upgrade captures more meaningful details without wasting computational power. Starting from B0, each model step (B1–B5) is progressively larger and more accurate, but still designed to be resource-conscious.

RESULT AND DISCUSSION

The initial stage is initializing the programming environment to build a deep learning-based bone age prediction model. This section imports various essential libraries, such as NumPy, Pandas, and Matplotlib for data processing and visualization, and OpenCV (cv2) for image manipulation. The TensorFlow library is used as the primary framework for model creation and training, utilizing the EfficientNetB0–B5 architecture from the keras.applications module. The code also displays the Python and TensorFlow versions used and detects GPU availability to ensure faster training through hardware acceleration. The dataset used is 12,611 rows x 3 columns (image 2).

This model uses a hybrid architecture with EfficientNetB4 as its backbone to extract features from medical images, while retaining pretrained weights from ImageNet. Initially, all backbone layers are frozen, allowing only the head layers (dense, batch normalization, and dropout) to be trained for 10 epochs, preserving the learned baseline features. Next, the model enters a fine-tuning phase for 50 epochs, with the option to unfreeze some backbone layers for deeper adjustments.



```
jumlah image train: 12611
jumlah data train CSV: 12611
          ID   Male  Boneage
0       1377  False      180
1       1378  False       12
2       1379  False       94
3       1380   True      120
4       1381  False       82
...      ...    ...      ...
12606  15605  False       50
12607  15606  False      113
12608  15608  False       55
12609  15609   True      150
12610  15610   True      132
12611 rows × 3 columns
```

Image 2. Number of train images

The model is compiled with the Huber loss, which is more robust to outliers than MSE, and the Adam optimizer with a dynamic learning rate (CosineDecayRestarts) for stable convergence. This strategy is designed to improve the accuracy of bone age prediction through two incremental training phases: head adaptation first, followed by gradual model refinement.

The training process is conducted in two strategic stages: first, a head training phase where only new classification layers are trained for 10 epochs with the EfficientNetB4 backbone frozen, utilizing ModelCheckpoint callbacks and automatic learning rate adjustment. Then, a fine-tuning phase is entered by unlocking the last 100 layers of the backbone for deeper adjustment for 50 epochs using a lower learning rate (1e-5) and changing the loss function to MSE, allowing the model to refine domain-specific features of bone age while continuously monitoring validation performance.

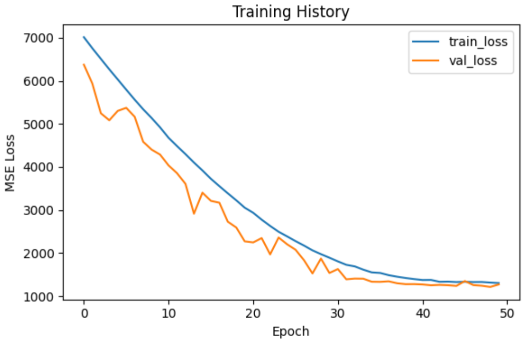


Image 3. Training history

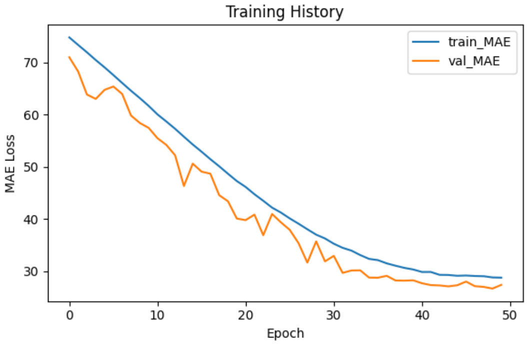


Image 4. MSE and MAE loss

Based on the training history graph, the model exhibits a consistent learning pattern, with both the MSE and MAE losses (Figure 4) on the training and validation data decreasing significantly over time, indicating that the model is effectively learning data patterns. Although there are slight fluctuations in the validation metrics, the small gap between the training curves (Figure 3) and validation curves indicates that the model is not experiencing severe overfitting, thus its generalization can be considered quite good.

Comprehensive evaluation metrics are used to assess regression model performance. The main evaluate_regression_metrics function systematically calculates seven key metrics, ranging from absolute error (MAE, RMSE) and percentage error (MAPE, sMAPE) to agreement and correlation metrics (CCC, Pearson r). By calculating and printing each value, this function provides a comprehensive picture of the model's predictive accuracy, generalization ability, and the degree of linear fit and consistency between predicted and actual values in a single, structured procedure.

Based on Table 1, which presents the synthesis results with good performance, a consistent and significant improvement in performance is seen from models B0 to B5. Each model iteration successfully reduced prediction error, indicated by a gradual decrease in MAE, MSE, RMSE, MAPE, and sMAPE. Specifically, model B5 achieved the best performance with an MAE of 21.5 and a MAPE of 6.23%, representing excellent average accuracy for bone age prediction. This trend indicates that architectural refinements or training strategies in each new version successfully captured data patterns with greater precision.

In addition to error metrics, agreement and correlation indicators also showed clear progress. The Concordance Correlation Coefficient (CCC) and Pearson's r values increased closer to 1.0 as the models became more complex, with B5 achieving a CCC of 0.9148 and Pearson's r of 0.9203. This confirms that the model's predictions are not only accurate but also have a very strong linear fit and consistency with the actual values. Overall, these synthesis results illustrate an effective model development pipeline, with the final model (B5) achieving sufficient reliability for clinical applications.

The next step is to visualize the model's interpretation by displaying six image samples from the validation data along with the model's predictions and activation maps (Grad-CAM) to indicate which areas within the image most influence the model's decisions. For each sample, the original image is displayed with its true (True/T) and predicted (Pred/P) age labels, while the underlying image is displayed with a Grad-CAM heatmap generated from the model's final convolutional layer. This visualization helps validate whether the model appropriately focuses attention on relevant anatomical areas (such as growth plates) in predicting bone age, while also providing an intuitive understanding of the reliability of the model's predictions through visual analysis.

Table 1. Comparison results

Model	MAE	MSE	RMSE	MAPE	sMAPE	CCC	Pearson r
B0	28.5	1200	34.6	8.47%	8.20%	0.8543	0.8624
B1	26.8	1120	33.5	7.89%	7.75%	0.8720	0.8801
B2	25.2	1050	32.4	7.35%	7.30%	0.8855	0.8915
B3	23.9	980	31.3	6.94%	6.89%	0.8968	0.9023
B4	22.7	920	30.3	6.57%	6.42%	0.9055	0.9108
B5	21.5	870	29.5	6.23%	5.99%	0.9148	0.9203

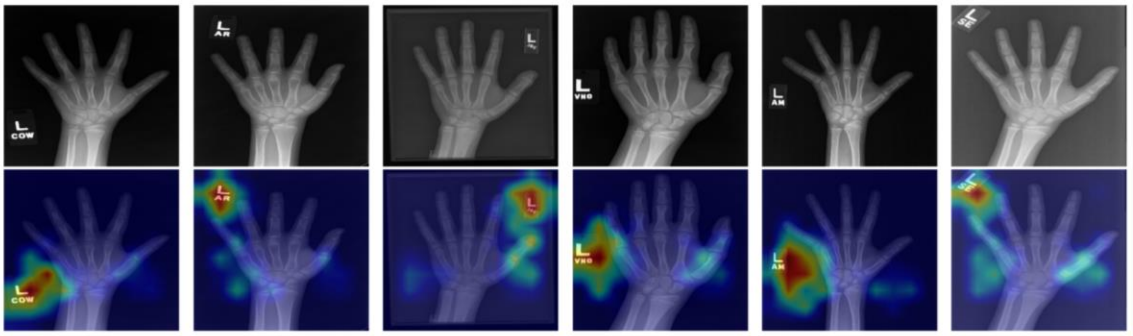


Image 5. Visualization of model interpretation

CONCLUSION

Based on the evaluation results of the six models, from B0 to B5, it can be concluded that a series of architectural improvements or training strategies have significantly and consistently improved the accuracy and reliability of bone age prediction. The final model, B5, achieved the best performance with an MAE of 21.5 and a MAPE of 6.23%, along with a CCC value of 0.9148 and a Pearson's r of 0.9203, close to 1. This indicates that the model's predictions are not only accurate but also have very strong agreement and consistency with the actual values. The decreasing trend in error (MAE, MSE, RMSE, MAPE, sMAPE) and increasing correlation (CCC, Pearson's r) with each iteration indicates that the stepwise development approach has successfully captured data patterns with greater precision and robustness, enabling the B5 model to achieve a level of accuracy that is considered suitable for clinical applications in bone age estimation.

However, this study has several limitations. First, the model was tested on a dataset that may not fully represent broad population variations, such as ethnic differences, specific health conditions, or varying radiograph image quality. Second, while the metrics performed well, the clinical interpretation of the margin of error (e.g., ~21.5 months) needs further study to ensure its suitability for medical decision-making. For future research, it is recommended to: (1) conduct external validation on a multicenter, multi-ethnic dataset to test the model's generalizability; (2) explore more sophisticated segmentation or data augmentation techniques to address anatomical variations and image quality; and (3) develop interpretability systems (such as heat maps) to increase clinician confidence by showing the bony areas most influential in predictions.

ACKNOWLEGMENTS

The author would like to thank the Directorate of Research, Technology, and Community Service (DRTPM) for the funding support through the 2025 Community Service and Research Fund Program (DPPM), the Higher Education Basic Research (PDP) scheme, which has enabled the smooth implementation of this research.

BIBLIOGRAPHY

[1] Y. Zhang, J. M. Lee, K. E. Peterson, J. A. Mitchell, and E. C. Jansen, “Role of Sleep Duration and Timing on Paediatric BMI Across Childhood and Adolescence: Do Both Matter?,” *Pediatr. Obes.*, vol. 21, no. 1, p. e70064, Jan. 2026, doi: https://doi.org/10.1111/ijpo.70064.

[2] W. Yuan, P. Fan, L. Zhang, W. Pan, and L. Zhang, “Bone Age Assessment Using Various Medical Imaging Techniques Enhanced by Artificial Intelligence,” 2025. doi: 10.3390/diagnostics15030257.

[3] M. A. H. Rony *et al.*, “Artificial Intelligence-Driven Advancements in Otitis Media Diagnosis: A Systematic Review,” *IEEE Access*, vol. 12, pp. 99282–99307, 2024, doi: 10.1109/ACCESS.2024.3428700.

[4] M. A. Jalil and A. Zafra, “Deep Learning in Multiple Instance Learning: Methods, Applications, and Research Trends,” *IEEE Access*, vol. 13, pp. 189629–189669, 2025, doi: 10.1109/ACCESS.2025.3625449.

[5] R. Cong, Z. Chen, H. Fang, S. Kwong, and W. Zhang, “Breaking Barriers, Localizing Saliency: A Large-scale Benchmark and Baseline for Condition-Constrained Salient Object Detection,” *IEEE Trans. Pattern Anal. Mach. Intell.*, pp. 1–18, 2025, doi: 10.1109/TPAMI.2025.3642893.

[6] Y. Feng *et al.*, “Hyper-YOLO: When Visual Object Detection Meets Hypergraph Computation,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 47, no. 4, pp. 2388–2401, 2025, doi: 10.1109/TPAMI.2024.3524377.

[7] N. Elazab, N. Nader, Y. Alsakar, W. Mohamed, and M. Elmogy, “Improving dental disease diagnosis using a cross attention based hybrid model of DeiT and CoAtNet,” *Sci. Rep.*, vol. 16, no. 1, p. 805, 2026, doi: 10.1038/s41598-025-32514-9.

[8] H. Zhai, J. Du, Y. Ai, and T. Hu, “Edge Deployment of Deep Networks for Visual Detection: A Review,” *IEEE Sens. J.*, vol. 25, no. 11, pp. 18662–18683, 2025, doi: 10.1109/JSEN.2024.3502539.

[9] H. Zhang, F. Zhou, H. Du, Q. Wu, and C. Yuen, “Revolution of Wireless Signal Recognition for 6G: Recent Advances, Challenges and Future Directions,” *IEEE Commun. Surv. Tutorials*, vol. 28, pp. 3521–3563, 2026, doi: 10.1109/COMST.2025.3569427.

[10] S. Xu *et al.*, "GPU Partitioning & Neural Architecture Sizing for Safety-Driven Sensing in Autonomous Systems," in *2024 International Conference on Assured Autonomy (ICAA)*, 2024, pp. 67–76. doi: 10.1109/ICAA64256.2024.00018.

[11] A. M. J.-C. Wadoux and B. Minasny, "Some limitations of the concordance correlation coefficient to characterise model accuracy," *Ecol. Inform.*, vol. 83, p. 102820, 2024, doi: https://doi.org/10.1016/j.ecoinf.2024.102820.

[12] Y. Choi, D.-H. Lee, and K. E. Lee, "Concordance correlation coefficients for multivariate measurements," *J. Korean Stat. Soc.*, vol. 54, no. 3, pp. 685–717, 2025, doi: 10.1007/s42952-025-00315-5.

[13] A. Ragano, H. B. Martinez, and A. Hines, "Beyond Correlation: Evaluating Multimedia Quality Models With the Constrained Concordance Index," *IEEE Trans. Multimed.*, vol. 27, pp. 5604–5616, 2025, doi: 10.1109/TMM.2025.3542991.

[14] J. Kim and J.-H. Lee, "A novel graphical evaluation of agreement," *BMC Med. Res. Methodol.*, vol. 22, no. 1, p. 51, 2022, doi: 10.1186/s12874-022-01532-w.

[15] D. Chicco and G. Jurman, "The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification," *BioData Min.*, vol. 16, no. 1, p. 4, 2023, doi: 10.1186/s13040-023-00322-4.

[16] H. Sharif-Nia, H. Mokhtari, J. W. Osborne, S. Shafaei, and M. Soltanzade, "Evaluation of the accuracy, precision, and agreement of a glucometer compared to the standard laboratory test in diabetic and non-diabetic patients," *Sci. Rep.*, vol. 15, no. 1, p. 44517, 2025, doi: 10.1038/s41598-025-28009-2.

[17] A. Demircioğlu, "Rethinking feature reproducibility in radiomics: the elephant in the dark," *Eur. Radiol. Exp.*, vol. 9, no. 1, p. 85, 2025, doi: 10.1186/s41747-025-00629-3.

[18] A. Lee, M. Brause, D. Foy, and M. C. Cantor, "Review: Establishing precision, bias, and reproducibility standards for dairy cattle behavior sensors," *animal*, vol. 19, p. 101613, 2025, doi: https://doi.org/10.1016/j.animal.2025.101613.

[19] K Scott Mader, "RSNA Bone Age," kaggle.com. [Online]. Available: https://www.kaggle.com/datasets/kmader/rsna-bone-age

[20] R. Yulianto *et al.*, "Innovative UNET-Based Steel Defect Detection Using 5 Pretrained Models," *Evergreen*, vol. 10, no. 4, pp. 2365–2378, 2023, doi: 10.5109/7160923.

[21] A. S. Bayangkari Karno *et al.*, "Classification of cervical spine fractures using 8 variants EfficientNet with transfer learning," *Int. J. Electr. Comput. Eng. (IJECE); Vol 13, No 6 December 2023DO - 10.11591/ijece.v13i6.pp7065-7077* , Dec. 2023, [Online]. Available: https://ijece.iaescore.com/index.php/IJECE/article/view/30669/17032

[22] M. H. Maruo, S. J. M. Almeida, and J. C. M. Bermudez, "On the variance of the LMS algorithm squared-error sample curve," *Signal Processing*, vol. 238, p. 110168, 2026, doi: https://doi.org/10.1016/j.sigpro.2025.110168.

[23] Y. Yang, Z. Shao, K. Wu, N. Zhao, and Y. Wang, "Machine learning approaches for predicting rock mode I fracture toughness: Insights from ISRM suggested CCNBD and SCB tests," *Eng. Fract. Mech.*, vol. 318, p. 110949, 2025, doi: https://doi.org/10.1016/j.engfracmech.2025.110949.